

DOCUMENT RESUME

ED 050 144

TM 000 531

AUTHOR Roudabush, Glenn E.; Green, Donald Foss
TITLE Some Reliability Problems in a Criterion-Referenced Test.
INSTITUTION McGraw-Hill Book Co., Monterey, Calif.
PUB DATE Feb 71
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE MF-\$0.65 HC-\$3.29
DESCRIPTORS *Behavioral Objectives, Correlation, *Criterion Referenced Tests, Diagnostic Tests, *Mathematics, Norms, Pilot Projects, Prediction, *Test Construction, Testing Problems, *Test Reliability, Test Validity

IDENTIFIERS PMI, *Prescriptive Mathematics Inventory

ABSTRACT

This paper describes the development of a criterion-referenced test. The Prescriptive Mathematics Inventory (PMI) was developed to measure 400 stated behavioral objectives. The test consists of three overlapping levels with the objectives chosen to cover 90 to 95 per cent of the mathematics curriculum nominally taught in grades 4 through 8. Each of the 400 items was designed to measure one of the 400 objectives, the criterion of mastery of each objective being that the student find the correct answer to the corresponding item. Two types of errors can be made when responding to each item: (1) by marking the correct answer when in fact the criterion behavior has not been mastered, and (2) by missing the item even though the criterion behavior has been mastered. To cope with these two types of errors the PMI provides a unique answer grid for the recording of answers to items. The PMI was tested with a group of fifth and seventh graders and it was noted that the students appeared to become bored or tired because of the length of the PMI. (CK)

ED050144

Some Reliability Problems in a Criterion-
Referenced Test

Glenn E. Roudabush and Donald Ross Green

CTB/McGraw-Hill

U.S. DEPARTMENT OF HEALTH, EDUCATION
& WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED
EXACTLY AS RECEIVED FROM THE PERSON OR
ORGANIZATION ORIGINATING IT. POINTS OF
VIEW OR OPINIONS STATED DO NOT NECES-
SARILY REPRESENT OFFICIAL OFFICE OF EDU-
CATION POSITION OR POLICY.

A Paper Presented at the American Educational
Research Association Meetings in New York
February 4, 1971

TM 000 531

Some Reliability Problems in a Criterion-Referenced Test

Any criterion-referenced test will be based on some set of explicit, behaviorally stated objectives -- the criteria against which an examinee's performance on the test is to be evaluated. The intent of the test is to show mastery or non-mastery of these objectives in an absolute sense, that is, not relative to any other person's performance. A norm-referenced test, on the other hand, is intended to show where an examinee stands with respect to some group -- his peers, a reference group, or some sampled standardization group. The use of a criterion-referenced test presupposes specific knowledge of the terminal behaviors desirable in the examinee; it also presupposes that this inventory of behaviors be comprehensive for the domain being measured. Further, a criterion-referenced test is generally intended to be relevant to classroom instruction. The items making up the test, then, should be sensitive to instruction and not unduly influenced by other variables. Any test of cognitive achievement will, in most circumstances, be highly correlated with tests of general intelligence, but this correlation should come about through an increased ability to profit from instruction and not through a direct relationship with intelligence test items.

Mager (1962) describes three parts of a well-stated objective: (1) that the desired outcome be stated in terms of observable behavior, (2) that the conditions under which the behavior is to be displayed be specified, and (3) that the criterion of acceptable performance of the behavior be stated. An example of such an objective in the math area is the following: "Given ten problems in the addition of two like fractions without regrouping,

the student will be able to find and record the correct answer to at least eight of them using paper and pencil." This statement seems quite straightforward. There is little ambiguity in it. Although the form of presentation of the problems is not specified, still one can easily imagine a student at his desk, pencil in hand, working diligently to get at least eight right. However, in this objective, as in all others, there are at least three implicit subjective judgements: first, that the behavior has value; second, that eight out of ten correct solutions represents mastery; and third, that the ten presented problems adequately represent the desired behavior domain. Stating instructional objectives in behavioral terms does not eliminate subjectivity. It does make our judgements public and, therefore, open to scrutiny and criticism.

There is general consensus that knowing some mathematics is of value to every student. How much mathematics and at what level of proficiency is less clear. It is probably important for every student to learn to add $1/5$ to $2/5$ and get $3/5$. Should every student also be able to add $1327/32475$ to $1124/32475$? Or, to get away from this simple example, should every student be able to find the cartesian product of two sets? In constructing criterion-referenced tests for general use, as we are doing at CTB/McGraw-Hill, the judgements of the value of the objectives to be measured by the tests are both difficult and critical, even in a discipline as well structured as mathematics. Initially, our approach has been to avoid theory, our own or anyone else's, and to look for consensus. Our procedure is to cull specific objectives from the text materials most widely used in the schools, collate those from each source into a single list, classify them into broader

objectives, and discover the categorical and hierarchal structure inherent in the objectives. Although this arduous and time-consuming process requires considerable dedication to completeness and accuracy, we believe it lends validity to the comprehensive and representative objectives measured by our tests.

A test item is a sample of behavior. This is as true of a criterion-referenced test as it is of a norm-referenced test, except that the behavior being sampled is a restricted and specifically defined criterion behavior of interest. The user of a criterion-referenced test wants to know in some absolute sense which criterion behaviors, that is, objectives, of those represented in the test the student has mastered and which he has not mastered, so that the student's further study can be directed toward those objectives of importance to him that he has not yet mastered. If the test is intended to be comprehensive for a discipline and if it covers material generally taught over a period of years, as is true of our math and reading tests, then the number of objectives that must be represented in the test becomes large. Unless the test is to become unmanageably long, very few items can be devoted to any one specific behavior. The number of items measuring each objective in the test and the criterion of mastery are crucial problems. If the test is to serve its function of giving specific diagnostic information about an individual student, then the information provided must be accurate and reliable.

A criterion-referenced test can be thought of as a large number of very short tests packaged together. Looked at in this way, the problem becomes a standard reliability problem, the solution for which is to find

statistically the minimum number of items required to measure each objective at some arbitrary acceptable level of reliability. Such an approach will lead to one of three unacceptable conclusions: (1) the test will be exceedingly long and therefore prohibitively expensive and time-consuming to administer and use; (2) the test will measure few objectives either by sampling objectives throughout the curriculum and therefore not being comprehensive, or by combining objectives and therefore losing diagnostic specificity; or (3) the test will consist of many levels, each covering only a small segment of the curriculum, and will again become prohibitively expensive and time-consuming to use. We suggest there is an alternative to this formulation of the reliability problem.

In our Prescriptive Mathematics Inventory (PMI), which is a criterion-referenced math test written by John Gessel, about 400 objectives are measured. The test consists of three overlapping levels that cover 90 to 95 per cent of the mathematics curriculum nominally taught in grades 4 through 8. Because of the large number of specific instructional objectives represented in the test, only one item is used to measure each behavior. Since there is only a single item measuring each objective in the test, the criterion of mastery of each objective is simply that the student find the correct answer to the corresponding item. The implicit assumption is that if the student gets an item correct he will be able to solve similar problems consistently without error and that if he misses an item he will consistently err in attempting similar items. This seems not to be a reasonable assumption, particularly if one looks upon the instrument as a collection of one-item tests.

In answering a test item, an examinee can err in two ways. First, if he has not mastered the criterion behavior, he can err by marking the correct answer. Second, if he has, in fact, mastered the criterion behavior, he can err by missing the item. Either kind of error is equally misleading when one is seeking specific diagnostic information about an individual student. The first kind of error occurs most frequently when there is a restriction placed on the number of alternative responses possible, as in multiple-choice or true-false tests. This kind of error can be counteracted by increasing the number of alternative responses available to the examinee. The second kind of error is more difficult to counteract since we have excluded the possibility of summing performance over a number of items that measure the same criterion behavior. The examinee is allowed no margin for clerical errors or other lapses.

In the PMI the first of these two kinds of errors, that of marking the correct answer even though the criterion behavior has not been mastered, is counteracted in the following way. The PMI provides for machine scoring of constructed responses by using a unique answer grid for the recording of answers to items. Two sample items are shown in Figure 1. Notice that the examinee first computes his answer, writes it in the spaces provided at the top of the grid, and then marks the grid. There are a hundred thousand ways to mark the first grid, only one of which is correct, and a thousand ways to mark the second grid. In the absence of any knowledge on the part of the examinee, the probability that he would mark the correct answer by chance is negligible. Some information is conveyed by the grid; however, in each case the grid is designed to fit the required answer exactly. The greatest

21 <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 20%; border: 1px solid black; text-align: center; padding: 2px;">1</td> <td style="border: 1px solid black; text-align: center; padding: 2px;">2/9</td> </tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">0</td><td style="border: 1px solid black; text-align: center; padding: 2px;">0 / 0</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">1</td><td style="border: 1px solid black; text-align: center; padding: 2px;">1 / 1</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">2</td><td style="border: 1px solid black; text-align: center; padding: 2px;">2 / 2</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">3</td><td style="border: 1px solid black; text-align: center; padding: 2px;">3 / 3</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">4</td><td style="border: 1px solid black; text-align: center; padding: 2px;">4 / 4</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">5</td><td style="border: 1px solid black; text-align: center; padding: 2px;">5 / 5</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">6</td><td style="border: 1px solid black; text-align: center; padding: 2px;">6 / 6</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">7</td><td style="border: 1px solid black; text-align: center; padding: 2px;">7 / 7</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">8</td><td style="border: 1px solid black; text-align: center; padding: 2px;">8 / 8</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">9</td><td style="border: 1px solid black; text-align: center; padding: 2px;">9 / 9</td></tr> </table>	1	2/9	0	0 / 0	1	1 / 1	2	2 / 2	3	3 / 3	4	4 / 4	5	5 / 5	6	6 / 6	7	7 / 7	8	8 / 8	9	9 / 9	7 <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 20%; border: 1px solid black; text-align: center; padding: 2px;">8</td> <td style="border: 1px solid black; text-align: center; padding: 2px;">7</td> <td style="border: 1px solid black; text-align: center; padding: 2px;">9</td> <td style="border: 1px solid black; text-align: center; padding: 2px;">4</td> <td style="border: 1px solid black; text-align: center; padding: 2px;">1</td> </tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">0</td><td style="border: 1px solid black; text-align: center; padding: 2px;">0</td><td style="border: 1px solid black; text-align: center; padding: 2px;">0</td><td style="border: 1px solid black; text-align: center; padding: 2px;">0</td><td style="border: 1px solid black; text-align: center; padding: 2px;">0</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">1</td><td style="border: 1px solid black; text-align: center; padding: 2px;">1</td><td style="border: 1px solid black; text-align: center; padding: 2px;">1</td><td style="border: 1px solid black; text-align: center; padding: 2px;">1</td><td style="border: 1px solid black; text-align: center; padding: 2px;">1</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">2</td><td style="border: 1px solid black; text-align: center; padding: 2px;">2</td><td style="border: 1px solid black; text-align: center; padding: 2px;">2</td><td style="border: 1px solid black; text-align: center; padding: 2px;">2</td><td style="border: 1px solid black; text-align: center; padding: 2px;">2</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">3</td><td style="border: 1px solid black; text-align: center; padding: 2px;">3</td><td style="border: 1px solid black; text-align: center; padding: 2px;">3</td><td style="border: 1px solid black; text-align: center; padding: 2px;">3</td><td style="border: 1px solid black; text-align: center; padding: 2px;">3</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">4</td><td style="border: 1px solid black; text-align: center; padding: 2px;">4</td><td style="border: 1px solid black; text-align: center; padding: 2px;">4</td><td style="border: 1px solid black; text-align: center; padding: 2px;">4</td><td style="border: 1px solid black; text-align: center; padding: 2px;">4</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">5</td><td style="border: 1px solid black; text-align: center; padding: 2px;">5</td><td style="border: 1px solid black; text-align: center; padding: 2px;">5</td><td style="border: 1px solid black; text-align: center; padding: 2px;">5</td><td style="border: 1px solid black; text-align: center; padding: 2px;">5</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">6</td><td style="border: 1px solid black; text-align: center; padding: 2px;">6</td><td style="border: 1px solid black; text-align: center; padding: 2px;">6</td><td style="border: 1px solid black; text-align: center; padding: 2px;">6</td><td style="border: 1px solid black; text-align: center; padding: 2px;">6</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">7</td><td style="border: 1px solid black; text-align: center; padding: 2px;">7</td><td style="border: 1px solid black; text-align: center; padding: 2px;">7</td><td style="border: 1px solid black; text-align: center; padding: 2px;">7</td><td style="border: 1px solid black; text-align: center; padding: 2px;">7</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">8</td><td style="border: 1px solid black; text-align: center; padding: 2px;">8</td><td style="border: 1px solid black; text-align: center; padding: 2px;">8</td><td style="border: 1px solid black; text-align: center; padding: 2px;">8</td><td style="border: 1px solid black; text-align: center; padding: 2px;">8</td></tr> <tr><td style="border: 1px solid black; text-align: center; padding: 2px;">9</td><td style="border: 1px solid black; text-align: center; padding: 2px;">9</td><td style="border: 1px solid black; text-align: center; padding: 2px;">9</td><td style="border: 1px solid black; text-align: center; padding: 2px;">9</td><td style="border: 1px solid black; text-align: center; padding: 2px;">9</td></tr> </table>	8	7	9	4	1	0	0	0	0	0	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6	7	7	7	7	7	8	8	8	8	8	9	9	9	9	9	21 Add: $\frac{7}{9} + \frac{4}{9} = \square$ $\frac{11}{9} = 1\frac{2}{9}$	7 Add: $\begin{array}{r} 38,296 \\ + 49,645 \\ \hline 87,941 \end{array}$
1	2/9																																																																															
0	0 / 0																																																																															
1	1 / 1																																																																															
2	2 / 2																																																																															
3	3 / 3																																																																															
4	4 / 4																																																																															
5	5 / 5																																																																															
6	6 / 6																																																																															
7	7 / 7																																																																															
8	8 / 8																																																																															
9	9 / 9																																																																															
8	7	9	4	1																																																																												
0	0	0	0	0																																																																												
1	1	1	1	1																																																																												
2	2	2	2	2																																																																												
3	3	3	3	3																																																																												
4	4	4	4	4																																																																												
5	5	5	5	5																																																																												
6	6	6	6	6																																																																												
7	7	7	7	7																																																																												
8	8	8	8	8																																																																												
9	9	9	9	9																																																																												

Figure 1. Sample addition items from the PMI, Level B

probability of obtaining a correct answer by chance occurs with those items requiring only a single digit response. For those items there is a probability of 1/10 that the correct answer would be marked by chance.

The second kind of error, that of missing the item even though the criterion behavior has been mastered, poses a more difficult problem. For any given objective measured by a single item in the PMI, we assume that additional information on that objective is tapped by other test items written to measure different, but related objectives. We are attempting to get at this additional information in several different ways. The first of these concerns the apriori relationships between objectives (items) in the test. The PMI objectives have been logically organized in a "Diagnostic Matrix" that provides a visual display of all the objectives measured by a given level of the test. A portion of the Level B matrix is shown in Figure 2. This section of the matrix concerns operations on the number line and the operation of addition. The numbers in

OPERATIONS AND THEIR PROPERTIES		WHOLE NUMBERS				POSITIVE FRACTIONS				DECIMAL FRACTIONS			
NUMBER LINE		1	2	3		6	7	8	9				
ADDITION	NO-REGROUPING				IDENTICAL FRACTIONS	14	EQUIVALENT			29			
					EQU. BALANCE	21	10THS			30			
					UNLIKE FRACTIONS	22	100THS			31			
					LIKE	23	1000THS			32			
					UNLIKE	24							
	REGROUPING				MIXED	25							
		1-4 DIGIT	18		LIKE	26	10THS			34			
		5-8 DIGIT	19		UNLIKE	27	100THS			35			
		COLUMN	20		MIXED	28	1000THS			36			

Figure 2. Section of the PMI Diagnostic Matrix, Level B

the small squares on the right are item reference numbers. The first seven items in the test are shown opposite the label "number line" and are all number line problems. The first of these is a simple concept question to determine whether or not the student can read a number off the number line. The second is a number line problem involving addition of whole numbers, the third subtraction of whole numbers, and so on.

In addition to the items shown in Figure 2, the complete matrix shows items for subtraction, multiplication, division, place value, expanded notation, mathematical sentences, measurement, geometry, etc., -- well covering the range of mathematics taught nominally in grades 5 through 7. Two similar matrices have been developed corresponding to Levels A and C of the test. Each item in the matrix (and in the test) represents a specific instructional objective in mathematics which can be stated in explicit behavioral terms.

As an example, consider the objective, "The student will be able to add like fractions without regrouping," represented on the Diagnostic Matrix as item 23. This is the only item representing that objective; however, items 24 through 28 differ only in the complexity of the operations of adding fractions. Items 23 through 28 are arranged in a hierarchal order with respect to this complexity.

The items themselves are as follows:

$$23. \quad \frac{1}{9} + \frac{2}{9} + \frac{5}{9} = \square$$

$$24. \quad \frac{1}{4} + \frac{3}{8} = \square$$

$$25. \quad 3\frac{1}{4} + 5\frac{1}{8} = \square$$

$$26. \quad \frac{7}{9} + \frac{4}{9} = \square$$

$$27. \quad \frac{3}{4} + \frac{5}{8} + \frac{1}{2} = \square$$

$$28. \quad 2\frac{5}{8} + 7\frac{3}{4} = \square$$

There are 64 different patterns of right/wrong scores possible for just these six items. Most of them are unlikely and would be difficult to interpret. Others could be quite informative. For example, if a student got items 24 and 25 correct, but missed items 23, 26, 27, and 28 one might provisionally conclude that the student has difficulty regrouping fractions and that he missed item 23 for some irrelevant reason. Since he correctly worked items 24 and 25, which are more complex than, but similar to item 23, he can, in fact, "Add like fractions without regrouping." In addition to the six items discussed here, there are six other items in the Level B test involving addition of fractions.

A thoroughgoing analysis of all the apriori relationships among all the items in the PMI would be voluminous and impossible to use. We are not suggesting that a classroom teacher be burdened with that number of considerations in diagnosing mathematics difficulties for each of his students. The Diagnostic

Matrix, however, presents aspects of these relationships in an immediate visual display. One of the student reports from the machine scoring of the PMI is a complete Diagnostic Matrix for each student with a correct (+) or incorrect (-) indication in place of the item reference number. With a little practice, a classroom teacher should be able to easily spot areas of deficiency for each of his students and for his class as a whole.

A second approach to the problem of single item reliability involves the empirical relationships between PMI items and scores on subsets of items. Last spring, a small pilot validity study was conducted for the PMI along with the item tryout, and additional data will be collected this spring. The pilot study involved the administration of the PMI and a criterion test to the same students. To construct the criterion tests, we first selected 40 items from the PMI which represent the range of major areas in mathematics covered by the test. For 12 of these 40 items, we wrote 20 alternate forms of the items. For the remaining 28, we wrote five alternate forms. These five-item or 20-item tests serve as criterion tests for the 40 selected PMI items.

We administered the appropriate level of the tryout version of the PMI along with the appropriate criterion tests of a sample to fifth grade and seventh grade students in a single school. Because of the length of the combined tests, however, very few students completed all of the items. In addition, it became apparent in examining the test booklets that many of the students were either too tired or too bored to do their best work. The results reported here, then must be considered tenuous.

In this initial study, we obtained the joint distribution of each criterion test score with its corresponding item from the PMI and the point-biserial correlation coefficient between them. Of the twelve 20-item criterion tests, nine have a sufficient number of cases to be indicative. For these nine tests, the average point-biserial correlation with the corresponding PMI item is .56 with a low of .33 for multiplication of decimal numbers and a high of .82 for addition of positive fractions. Of the twenty-eight 5-item criterion tests, 16 have a sufficient number of cases to be indicative. For these 16 tests, the average point-biserial correlation with the corresponding PMI item is .43 with a low of .07 for a probability item and a high of .80 for a problem concerning the greatest possible error of a physical measurement.

The wide variation in the size of these correlations is surprising, even considering the poor data and small number of cases upon which they are based. The extremely low correlation of .07 for the probability item with its criterion test can probably be accounted for by a wide variation in the item difficulties in the criterion test itself, indicating that the items making up the test are not alternate forms of the original item. Dropping this test out, however, still leaves an average point-biserial correlation of .46 with a low of .25 for subtraction of liquid measures for the 5-item tests. There is, of course, considerable variation in item difficulties for all the criterion tests, but this may primarily reflect the poor testing conditions. Still, the large differences in the correlations seem to suggest correspondingly large differences in the ability of particular items to accurately reflect mastery or non-mastery of the criterion behavior they are designed to measure. The data we plan to collect this spring should give us better evidence on this possibility.

- . A second part of our reliability studies will be undertaken after we have collected additional data. In this study, we will develop regression equations to predict item criterion test scores from subscores of the PMI itself. PMI subscores were computed by adding together item scores for groups of items that logically go together (such as the six addition items discussed above). Standard item analysis techniques were used to check these subscores for internal consistency. The KR20's for these subscores generally run in the high .80's, but some adjustments will be made in the items included in some of the scores. We will use all of the subscores of the PMI as predictors of the item criterion test scores for each of the 40 selected items. This will, in effect, use all of the available information in the test to estimate the student's ability to perform on each of the specific objectives measured by the item criterion tests. If this procedure gives us a much more reliable estimate of the student's mastery level of an objective, we may use the same method to develop regression equations for all of the items in the PMI, and then incorporate the obtained regression equations in the scoring program for the test.

REFERENCE

Mager, R.F., Preparing Instructional Objectives, San Francisco: Fearon Publishers, 1962.